AI, Attachment Theory and Simulating Secure Base Behaviour: Dr. Bowlby meet the Reverend Bayes.

Dean Petters[1] **Everett Waters**[2]

Abstract. In this paper we describe how information processing constructs originating in AI have become part of the Attachment Theory tool kit. We survey the early influence of AI within the theoretical framework that John Bowlby formed as the foundation of Attachment Theory between the 1950s and 1980s. We then review recent work which has built upon Bowlby's framework and is concerned with modelling and simulating attachment phenomena. We conclude by discussing some possible advantages that might arise from incorporating recent work on Bayesian arbitration into attachment control system models.

1 Introduction

John Bowlby formulated his Attachment Theory whilst working in a multidisciplinary team that included childcare professionals, psychoanalysts, ethologists and other researchers at the Tavistock Clinic for over thirty years after the Second World War (10; 21). Bowlby was interested in understanding issues such as: the separation distress exhibited by children when they or their mothers were absent due to the infant's or mother's hospitalization (9); the effect of early maternal deprivation on later development (4); and grief and mourning in infancy (5). From its inception Attachment Theory integrated concepts from academic fields as diverse as Ethology, AI and Psychoanalysis. One of Bowlby's primary goals was to replace Freud's drive theory with an attachment motivation theory rooted in modern scientific methods, empirically accessible, and better able to account for infant behavior's sensitivity to context. To accomplish this, he first turned to ethology and developed a framework that described the attachment system as an instinctive behaviour system (29). However, in his more mature theoretical work, Bowlby drew increasingly on control systems theory and on AI based representational constructs such as Internal Working Models and hierarchical planning.

Bowlbys Attachment Trilogy (6; 7; 8) contains his most developed behavioural descriptions and theoretical explanations for attachment phenomena. These accounts comprise a set of observable behaviours related to social and emotional attachment in animals and humans; and the cognitive mechanisms that give rise to these observable phenomena. Attachment Theory can therefore be presented as incorporating two theoretical components that can be termed the behavioural and cognitive components (21). Attachment Theory's behavioural component is valuable for researchers interested in modelling and simulation because the set of attachment behaviours that have been observed and recorded over the last half century provide data for numerous behavioural scenarios. These behavioural descriptions can be interpreted and adapted to act as standardised specifications of requirements allowing simulations to be focused in their evaluation (25; 26). Bowlby did not describe Attachment Theory's cognitive component in enough detail to be straighforwardly implemented in simulation without additional interpretation and development. What this can provide is a point of departure for contemporary AI theories. This means that simulation designers can use the information processing explanations for attachment phenomena that already exist within Attachment Theory as a starting point and inspiration when incorporating new AI structures and mechanisms.

2 The Behavioural Component of Bowlby's Attachment Theory

The offspring of many animals, including humans, show a tendency to gain and retain physical proximity to their main carer, usually their mothers. This is not surprising, as for many of these species the mother feeds her offspring, and infant animals have to approach to be fed. Some additional proximity seeking by the offspring to its mother may be due to secondary reinforcement effects that are driven by the primary reinforcement of gaining food. However, animal studies have shown proximity seeking to a main carer can be unrelated to any reinforcing effect of feeding. Research on imprinting in Geese (23) and research with infant monkeys and wire-frame mother substitutes (20) are amongst a number of studies that show animals can imprint or attach to individuals or objects that do not provide a source of food.

From the late 1930s Bowlby was undertaking human studies on attachment that can be seen as a parallel to research on attachment processes in animals. The key insight in interpreting the attachment behaviours linked to proximity seeking that Bowlby described is to see the attached infant or child as using their carer as a secure base from which to explore. When infants develop the ability to crawl, and later to walk, they can explore the world more effectively and the rate at which they acquire knowledge about the broader world accelerates. However, this new found ability to explore brings with it the potential for access to many more hazards. A naturalistic study of toddlers between ages 1 and 2 investigated how infants balance the opportunity for exploration and the security provided by their carers (2). In London parks, infant and carer pairs were observed without their awareness, in observation periods averaging 15 minutes' duration. This study found that most infants moved away from their carers to explore, but kept within a caregiver's line of sight, and periodically 'checked-in' by gaining the attention of their carers or by moving back to closer proximity.

Attachment development and individual differences have also

[1] Newman University College, Birmingham, UK, email: d.d.petters@cs.bham.ac.uk

[2] State University of New York at Stony Brook, email: everett.waters@sunysb.edu

been examined through the lens of a standardised laboratory procedure - the Strange Situation (1). This procedure is not an experiment where subjects are randomly assigned to different conditions in the laboratory. Rather, it presents all infants with the same controlled and replicable set of experiences. Nested within the normative trends that illustrate infant's sensitivity to context are several patterns of response reflecting the infant's confidence in the caregiver's response patterns.

A key research goal for the Strange Situation procedure was to demonstrate experimentally that infant attachment behaviour was sensitive to context - as opposed to rising and falling with the drive states of psychoanalytic theory (which would have a longer and more regular course and no obvious link to context). To capture infant responses to changes in context, the Strange Situation procedure consists of 8 three minutes episodes which are designed to activate, intensify or relax one-year-old's attachment behaviour in a moderate and controlled manner. The context changes that occur in the transitions between the eight episodes, and the and the infant's responses to these transitions, don't just help test a control systems approach against a drive state approach to explaining attachment. They also provide a valuable data-set for contemporary researchers interested in designing attachment behaviour simulations.

In the Strange Situation, the infant and carer enter the laboratory setting together, but then undergo a separation, when the carer leaves from the room, before a reunion in a subsequent episode. After the first reunion episode the infant also meets an unfamiliar 'stranger' in the laboratory, before a further separation. In each episode the infant's behaviour is carefully recorded from behind a two-way mirror. In the final episode the mother returns to her one-year-old infant after the infant has been left alone for three minutes in the unfamiliar setting. The infant's response in the reunion episodes correlates strongly with patterns of maternal behaviour and infant responses intensively observed throughout the previous year. An infant's responses to reunion in the Strange Situation can therefore act as a shorthand for the infant's home relationship with their carer (1). Normative behaviour patterns across episodes highlighted the infants' sensitivity to context. There were also marked individual differences which clustered into three distinct patterns, labelled: Secure (type B), Avoidant (type A), and Ambivalent (type C). More recent studies have categorised a fourth type of Disorganised pattern of attachment (type D) that is the least well characterised or understood and forms a very small proportion of infants in the general population (19, page 26).

Secure infants are the largest group and secure behaviour is the reference pattern against which the other classifications are evaluated. These infants respond to their mothers on reunion in the Strange Situation by approaching them in a positive manner. They then return to play and exploration in the room quickly. They received care at home which can be summarised as being consistently sensitive. In comparison with average levels across all groups, mothers of B type infants were observed at home being more emotionally expressive and provided less contact of an unpleasant nature; at home these infants were less angry and they cried less.

Avoidant infants respond to their carer on reunion in the Strange Situation by not seeking contact or avoiding their carer's gaze or avoiding physical contact with her. These children return quickly to play and exploration but do so with less concentration than secure children. Whilst playing they stay close to and keep an eye on their carer. It may seem that they are not distressed or anxious in the Strange Situation. However, research employing telemetered autonomic data and salivary hormone assays has demonstrate that, despite their relative lack of crying, avoidant infants are at least as stressed by the procedure as secure and resistant infants (19, page 193). In comparison with average levels across all groups, mothers of A type infants were observed at home being consistently less sensitive to infant signals and less skilled in holding the baby during routine care and interaction. However, in the reunion episodes these infants showed the least anger and crying.

Ambivalent infants respond to their carers on reunion in the Strange Situation by: not being comforted and being overly passive or showing anger towards their carers. These infants do not return quickly to exploration and play. They received care at home which can be summarised as being less sensitive and particularly inconsistent. In comparison with average levels across all groups, C type carers were observed at home being more emotionally expressive; they provided physical contact which was unpleasant at a level intermediate between A and B carers and left infants crying for longer durations; at home these infants were more angry, and they cried more.

Disorganised infants responses in the Strange Situation do not form a clear pattern. Displayed behaviours may be categorised as: contradictory, lacking direction, anomalous, dazed, apprehensive, or disoriented. These infants experience particularly unpredictable and inadequate home caregiving (19, pages 25- 26).

Strange Situation studies provide correlational results that link patterns observed in early carer and infant behaviour with patterns observed in later infant behaviour. Several lines of evidence support an interpretation of these patterns in terms of the infant adapting to their pattern of caregiving they receive, rather than due to innate carer and infant temperaments (28; 19, pages 53-80). Thus for simulation design, infant attachment behaviourat home can be viewed as training data and Strange Situation behaviouras test data. Just as attachment models need to accommodate both the normative findings across episodes as well as the individual differences, they should also take into account that attachment patterns are both sensitive to recent experience (1, pages 217-219) and yet tend to be quite stable across time (19, page 243). If repeated within 10-14 days, infants seem to recognize the test situation, show more distress, more proximity and contact seeking, and the A,B,C patterns are less distinct. And yet the classifications can be quite stable across longer intervals in infancy. The reliability and stability of individual differences in infant-mother attachment predict attachment related measures even in early adulthood (30). An individual's patterns of secure base behaviour therefore remain relatively stable across numerous developmental stages, a noteworthy finding considering the radical changes in an individual's cognitive machinery that occur across these transitions. Perhaps what is common for securely attached individuals across developmental stages is their expectations about their attachment figure's availability and responsiveness in acting as secure-bases. Though the form of these secure-base representations may vary from implicit representations in infancy to secure base scripts in adults (31).

3 The Cognitive Component of Bowlby's Attachment Theory: AI and other influences

3.1 Bowlby's conceptual journey

Bowlby published scholarly articles on attachment phenomena between the 1940s and 1980s. During this period the theoretical framework for Attachment Theory was refined and came to incorporate more sophisticated information processing concepts:

> *"The hypothesis proposed represents a development of that advanced by me in 1958. The principal change is due to better understanding of control theory and to recognition of the*

> *very sophisticated forms that behavioural systems controlling instinctive behaviour may take. In the present version of the hypothesis it is postulated that, at some stage in the development of the behavioural system responsible for attachment, proximity to mother becomes a set-goal. In the earlier version of the theory five patterns of behaviour - sucking, clinging, following, crying, and smiling - were described as contributing to attachment. In the new version these same five patterns are still held to be of great importance, but it is postulated that between the ages of about nine and eighteen months they usually become incorporated into far more sophisticated goal-corrected systems. These systems are so organised and activated that a child tends to be maintained in proximity to his mother."*
>
> *"The earlier version of the theory was described as a theory of component instinctual responses. The new version can be described as a control theory of attachment behaviour"* (6, page 180,)

Bowlby's new version of Attachment Theory shows the continued importance of secure base behaviour with an increasing role for mental representation. However, the new theory still includes a strong ethological influence. For example, Bowlby still presents the attachment system as an instinct to form bonds and as a system that is activated by species specific patterns of care. However, this new version also emphasises three additional components. As a control theory there is a greater focus on the attachment system as directed towards outcomes as set-goals to be achieved from a flexible behavioural repertoire (rather than a system that simply involves triggering preset responses). This control system is also described as possessing a sophisticated variety of algorithms, representations, and architectural detail, such as hierarchical structures. Also, the control system as proposed by Bowlby is not just preformed and waiting to be triggered or maturing without experience, but its rather constructed - through interaction between infant learning abilities and information available in the structure of the caregiving environment. With such information processing concepts as these at its core, attachment theory was well positioned to exploit advances in AI. (6; 26).

3.2 The attachment control system carries out species specific functions and is inherently motivated.

Although the control systems formulation was a major departure from Bowlby's early instinct theory, he retained his commitment to behavioral biology. The theoretical inheritance shouldn't be underplayed, as Hinde notes:

> *"The concept of a behavioural system is, in fact related to one meaning of the term instinct. [...] It has been used in a rather special sense by ethologists to refer to systems postulated as controlling a group of behaviour patterns that together serve to achieve a given biological end"* (Hinde 1983, page 57).

In animal ethology, behaviour systems are theorised as controlling behaviours such as mating, fighting and feeding. Each behaviour system carries out a species specific function, and has been selected for this function in the evolutionary past. Bowlby suggested that infants possess a somewhat similar species-specific behaviour system that lead to predictable outcomes which are likely to contribute to reproductive fitness. The behaviour systems that Bowlby linked to attachment behaviour in human infants are the attachment, fear, sociability and exploration systems (6). For Bowlby, behaviours resulting from the attachment behaviour system and the fear system have the predictable outcome of maintaining access and proximity to its primary carer - its secure base and haven of safety. The exploratory behaviour system activates behaviours that result in learning and the sociable system results in social interaction.

A second common thread in the evolution of Attachment Theory is that the behaviour systems most closely related to attachment are inherently motivated. Infants will work to experience exploration, socialisation and security because these outcomes can be considered primary drives. They are not activated as the by-product of any more fundamental process. This means that infants' interest in exploration is a primary motive, not derived from feeding or contact comfort (6). Actions such as running away, freezing and using a carer as a secure base are all behaviours that humans and other animals seem to instinctively 'know' how and when to do when faced with particular dangers. However, this does not mean that these behaviour patterns are not themselves constructed from more basic behavioural components, as we shall see in section 3.5.

3.3 The attachment control system uses a flexible repertoire of behaviours in pursuit of set-goals.

According to Bowlby, what defines the attachment control system is not a set behaviour repertoire but the outcomes that predictably follow from these behaviors (6). Similar behaviours may be produced by different behaviour systems. For example, behaviours such as locomotion can serve more than one system, such as the attachment and exploratory system. Also any given behaviour system may produce a wide range of differing behaviours. In the attachment system, if the infant's goal is to increase its proximity to a carer the infant may cry (which predictably brings the carer closer), or crawl towards the carer themselves. This is an example of behaviours within systems being interchangeable with other functionally equivalent behaviours.

In this formulation, the behaviors used to gain and maintain proximity range from subtle signals, such as gazing towards a carer, to overt signals, such as calling a carer, and active behaviours, such as locomotion. Exploratory behaviours range from locomotion to object manipulation and have the predictable outcome of improving the infant's ability to manipulate the external world.

3.4 The attachment control system involves a hierarchy of forms of information processing

In describing Behaviour Systems, Bowlby (6) invoked a hierarchy of information processing tools which include:

- Ethological concepts and mechanisms, such as Behaviour Systems, Reflex Actions and Fixed Action Patterns which can interact in complex ways by chaining and alternation;
- Concepts from Control Systems theory such as feedback and goal directed mechanisms;
- Concepts from AI and Cognitive Science such as Internal Working Models (IWM's) and hierarchical organisation and control of behaviour using complex representational forms such as hierarchical plans, and natural language.

Reflexes are behaviours with a highly stereotyped form. Once activated by a stimulus at a specific threshold they are ballistically carried to completion. Fixed action patterns are similar to reflexes because they are stereotyped but differ from reflexes because they are open to learning. The thresholds for activation and termination adapt according to the organism's state and past experiences. Fixed action patterns are also less ballistic, with for example, proprioceptive feedback during execution. Examples include grasping, crying and smiling. Reflexes predominate in the first few months after birth and fixed action patterns predominate from three months until the middle and end of the first year. The reflexes and fixed action patterns that infants perform may seem a very simple form of control but are highly effective in eliciting adult actions that benefit the infant. Different reflexes and fixed action patterns are coordinated together. These complex patterns produced by fixed action patterns can be mistaken for behaviours directed by goal corrected mechanisms because of the sensitive matching of response to stimuli.

The final stage in attachment development, emergence of what Bowlby called a goal-corrected partnership, commences from the middle to the end of the first year. When a mechanism is goal-corrected it is updated or retaken according to feedback on how well the goal has been satisfied. When several goal corrected steps are chained together, and each step must be completed before the next step is taken what has been formed is a plan. Bowlby considered simple plans, and more complex plan hierarchies. He discussed the substructures from which more complex goal corrected plans might be constructed. He also emphasised that, as with orders in a military operation, the attachment control systems involve producing high level plans that are then translated into lower level plans. However, within some limits the lower level subplans can be varied, leading to the flexibility in behaviour repertoire that has already been mentioned above in section 3.3.

Bowlby adapted the working models concept to a more restricted attachment specific concept of Internal Working Models (IWM's), which represent models of self and other in attachment relationships[3]. In this usage, IWM's are held to capture the relation-structure of attachment phenomena, not every aspect of reality but enough that the child can formulate and choose among. These include spatio-temporal causal relations among the events, actions, objects, goals and concepts represented. IWMs represent attachment related world knowledge and expectations about its caregiver's availability and responsiveness. These expectations are derived from the carer's past performance. IWM's of self and attachment figure develop in a complementary manner. An important challenge for current attachment research and theory is to specify in greater detail the cognitive architecture and content of IWMs at different ages. For example if the carer is responsive the self is valued.

Bowlby emphasised the importance of updating IWM's of self and environment:

> *"To be useful both working models must be kept up-to-date. As a rule this requires only a continuous feeding in of small modifications, usually a process so gradual that it is hardly noticeable. Occasionally, however, some major changes in environment or organism occur: we get married, have a baby, or receive promotion at work; or, less happpily, someone close to us departs or dies, a limb is lost, or sight fails. At those times radical changes of models are called for. Clinical evidence suggests that the necessary revisions of models are not always easy to achieve. Usually they are completed but only slowly, often they are done imperfectly, and sometimes not at all."*(6, page 82)

Reflecting on how IWMs are updated, Bowlby compared the IWM's of his Attachment Theory with the 'internal worlds' of traditional psychoanalytic theory. He pointed out that that both perspectives assign mental representation a central role in the origins of psychopathology. He notes that the pathological sequelae of separation and bereavement can be understood in terms of models that are partly out of date or full of inconsistencies and confusions (6, page 82).

According to Bowlby, natural language is the ultimate and most sophisticated way in which an individual can represent themself within their social environment. This form of representation has the benefit that *"instead of each one of us having to build his environmental and organismic models entirely for himself, he can draw on models built by others"* (6, - page 82)[4]. A benefit of the non-communicative aspect of language is that the possession of language allows more flexible and imaginative plans and subplans to be created.

3.5 The attachment control system is constructed and reconstructed in a developmental process throughout the lifespan.

Bowlby's view of instinctual behaviour changed significantly over the 30 years he devoted to attachment theory. In his final presentation of Attachment Theory this concept was entirely congruent with a developmental theory of control system construction. According to Bowlby, the "instinctual" behaviors evolution provides are not limited to simple reflexes. They can be complex, goal corrected, and socially relevant. As an example, Bowlby (6, page 44) points to the the falcon's ability to capture small birds in flight as a quintessential example of an instinct: the falcon's stoop is species specific and appears largely without opportunity of learning. However, Bowlby, emphasises that even if we don't understand how control systems of this kind come to develop, we shouldn't doubt that there is a developmental explanation that relies upon the Falcon interacting within the ordinary expectable environment for its species.

Bowlby suggests that there are multifarious processes whereby early appearing fragments of instinctive behaviour are integrated into later appearing complete sequences with their normal mature functional consequences. These integrative processes include restrictions appearing in the activating and terminating conditions of the behaviours, and behaviours becoming units in one or more chains. Bowlby also describes a third process:

> *"Yet another sort of process [of functional integration] is one that integrates a piece of behaviour within a causal hierarchy. This can occur following a change in the causal relation between a pattern of behaviour and the internal state of the animal.*
>
> *It might confidently be supposed that feeding behaviour would be most readily elicited when an animal is hungry, and the hungrier it is, it might be though, the more readily would the behaviour be elicited. This is by no means so, however, at least in the very young. For example, when a fledgling great tit starts to peck it is most likely to do so when it is* **not** *hungry: when it is hungry it begs food from its parents. Similarly, experiment suggests that sucking behaviour in puppies is at first independent of both hunger and food intake. Later in development pecking and sucking become elicited most readily in conditions of hunger and, by those means, are brought, together with other behaviour contributing to food-intake, within a system organised in terms of causal hierarchy."* (6, pages 159-160,)

[3] Kenneth Craik originally introduced the term 'working model'. However, as Bretherton (10) notes Craik's concept of Working Models came to Attachment Theory indirectly from the through the writings of the biologist J. Young (1964).

[4] compare with Dennett's description of Gregorian Minds (17, page 99)

Clearly this process of causal hierarchy formation in great tit chicks doesn't result in the sophisticated kinds of representational redescriptions characteristic of human social and cognitive development. However, Bowlby goes on to set out a framework for how the kinds of processes that integrate behaviours in chicks and puppies can have far reaching effects in humans:

> *"Because of a human's immense capacity to learn and to develop complex behavioural systems, it is usual for his instinctive behaviour to become incorporated into flexible behavioural sequences that vary from individual to individual. Thus once a human has had experience of reaching a consummatory situation the behaviour that leads to it is likely to become reorganised in terms of a set-goal and a plan hierarchy"* (pages 160, 6)

These developmental processes described by Bowlby involve intimate interaction between lower level processes, such as simple reflexive responses, and emerging higher level structures and mechanisms. New resources come online gradually. Integrating elements into a system depends on (i) biases in infant learning abilities and (ii) information/structure in the expectable caregiving environment. Absent the knowledge about the latter social/caregiving affordances, and the former biases in learning abilities are of no use. Adult cognitive abilities are radically different from an infant's, but humans do not undergo some kind of process of metamorphosis like cognitive-caterpillars changing to cognitive-butterflies, where a brand new architecture emerges suddenly. Rather, new resources are broken in slowly, and take shape in an environment where the ongoing behaviour patterns have been set by the older earlier maturing resources.

4 Updating the Cognitive Component: designing and implementing simulations

Despite Bowlby making a valuable contribution towards explaining attachment phenomena in information processing terms, his ideas are not set out as design specifications merely awaiting implementation. He may have referred to ideas in Control Systems Theory and AI as much to break the old Freudian paradigm than to specify in great detail a way forward to more involved modelling and simulation. However, even if this were the case, Bowlby's information processing framework can still inform simulation design, but it must be interpreted, and in some sense extrapolated.

A number of different attachment control architectures have been implemented as simulations. These include simulations of: approach and exploratory behaviour in naturalistic environments (3; 26, pages 51 - 78); infant agents adapting to the secure or insecure patterns of caregiving agents (26, pages 79 - 102); and infant and carer agents reproducing the different behaviour patterns observed in the Strange Situation Experiment (26, pages 103 -152).

Most of the examples above involve autonomous agent architectures which possess a generic family resemblance. Though they possess multiple components, they can be considered homogenous architectures (in contrast with hybrid architectures) because each component is a single generic type of reactive subsystem (the components being similar to those incorporated in Behaviour Based architectures in robot or autonomous agent simulations (32; 26, pages 62 - 69)). Each of the different component 'behaviours' correspond to one of the behaviour systems that Bowlby invoked as involved in secure-base behaviour (exploration, sociable, attachment-security and wariness). In these homogenous architectures, the highest activated behaviour can control the actions that are carried out by the agent. However, the decision processes is not strictly winner take all because two goal activators with high activation may mutually inhibit each other, leaving a less highly activated goal to direct behaviour, and give the impression of displacement activity. Key attachment phenomena in real infants can be explained by interaction among the same type of information processing components incorporated in these agent based simulations. For example, A type infants have experienced aversive responses at home, particularly in close proximity. The phenomenon of A type infants avoiding their carers in reunion episodes of the Strange Situation can therefore be explained by the postulation of reactive avoidance behaviour systems in these infants becoming more activated than other components in the emotionally aroused environment of reunion episodes.

Interesting response patterns can emerge in simulations with infant agents possessing relatively simple homogenous reactive architectures. For example, computational experiments demonstrate novel epigenetic trajectories for the development of Secure and Insecure attachment behaviour patterns (26; 27). A key parameter in these simulations is the infant agent's 'confidence' in its carer agent's responsiveness. This parameter is represented in the infant agent's architecture as a 'safe-range' separation distance between infant and carer agent that the infant will tolerate without any rise in activation for the attachment anxiety goal activator. When separation goes beyond the infant agent's 'safe range' the attachment anxiety goal activator starts to become activated, and therefore competes with other goal activators for control of setting the next action. The attachment goal activator will become deactivated by the proximity to the carer agent increasing, or by the infant gaining the carer agent's visual attention. Simulations have demonstrated that for certain infant agent architectures with certain initial parameters, the infant and carer agents can interact in a positive feedback process that results in moderate levels of infant confidence in the carer agents responsiveness becoming unstable over lengthy periods of interaction. This instability results in populations of identical infant-carer agent dyads with initially moderate settings for confidence and responsiveness bifurcating into Secure and Insecure subgroups. This is interesting because clustering of attachment groups is found in empirical results from Strange Situation studies (1; 27; 26, pages 79 - 102). Bifurcation occurs because the effect of very small random variations in movements and timings on the otherwise identical infant-carer agent interactions are greatly amplified by the positive feedback. These initially identical agents end up as two groups with no intermediate cases and therefore could be categorised as separate taxons (24; 26, pages 97 - 99).

A contrasting example of a more complex attachment control system that has been implemented is the hybrid architecture with reactive and deliberative subsystems shown in figure 1. This hybrid architecture possesses reactive behaviour based components (similar to those described above in the homogenous architectures above), that represent information implicitly. However, in this architecture these reactive subsystem interact with a higher level deliberative reasoning subsystem that represents goals and the state of the world explicitly. This means that this architecture can construct explicit plans, which are composed of valid operations from the current the state of the world towards desired goal states. Figure 1 shows the higher level deliberative subsystem can interact with the lower system by inhibiting a reactive prepotent response (denoted by the dashed line), when this response may result in an undesirable outcome. Relatively sophisticated processes of reasoning and inhibition such as this may be easily overwhelmed by lower level reactive processes for one year old infants in the emotionally aroused context of the Strange Situation. However, in older children and adults such reasoning about social and emotional goals may interact with reactive processes and both influence attachment behaviour patterns. However, this particular example of a hybrid architecture is too simple and inflexible to capture the richness of attachment responses in the older age groups.

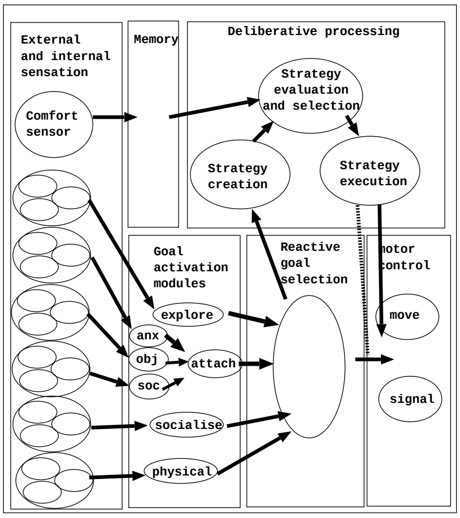

Figure 1. A Hybrid Design for the Attachment Control System. Deliberative mechanisms provide a secondary route to action, activated as a result of interrupts to reactive selection and arbitration and imposing different actions. The reactive goals are: exploration (explore); anxiety (anx); object wariness (obj); social wariness (soc); socialisation (socialise); and physical need (physical). The reactive goals of anxiety, object wariness, and social wariness are all combined into a single attachment goal (attach) before they are considered for selection. The comfort sensor measures contact pleasure and represents this information symbolically. Avoidance can occur when the deliberative subsystem inhibits approach because it has reasoned about the undesirability of close interaction.

Reactive architectures that lack the kinds of deliberative resources shown in figure 1 can simulate a wide range of naturalistic and laboratory attachment phenomena observed in infants. However, for older children and adults, deliberative subsystems more sophisticated than that presented in figure 1 would be expected to play some part in producing attachment behaviours. This raises a number of questions. What kinds of subsystems influence attachment behaviour at different ages? How will different subsystems interact? How will the empirical finding of continuity in attachment patterns be supported as higher level subsystems come 'on-line'? Lastly, how might new subsystems be constructed?

5 Future steps - Applying Bayes to Bowlby?

Dayan and colleagues (12; 15; 16; 22) have developed a theoretical framework that may help provide traction in answering some of these questions. This framework includes four action controllers, labelled: Pavlovian; Habitual; Model-based; and Episodic (13). These four controller possess complementary strengths and weaknesses and their operation is integrated together. In the context of attachment, a coherent behaviour pattern might be driven by more than one controller. For example, the initially observed behaviours in a Strange Situation episode might produced by the Pavlovian controller (whose responses are likely to be fastest). Other controllers may exert their influence later in the episode, but still be recognised as within the same attachment pattern response.

Recent evidence suggests how these controllers may be instantiated in specific brain areas (12; 13; 16; 22). In addition, the responses of the Habitual, Model-based and Episodic controllers are considered to be mediated by a decision principle based upon uncertainty (12; 13; 22) . Each of these three controllers incorporates a measure of its own uncertainty in its likely performance. The least uncertain controller is then chosen to direct the next action. This Bayesian decision mechanism may account for the Disorganised (type D) pattern of infant response in the Strange Situation. This pattern of response arises when infants receive unpredictable maternal care. One can imagine, for example, a Bayesian arbitration system producing "disorganized" behaviour in the face of unpredictable maternal care. In this case no controller would reach the level of certainty needed to direct action with coherence and consistency.

5.1 The Pavlovian controller

This controller performs routine and reflex behaviours such as fleeing from danger (15). Unlike the other three controllers the Pavlovian controller does not rely upon the individual learning from experience to adapt how it responds. Rather this controller possesses a set of responses that are tied to triggers that have been set in an evolutionary process. The reflexes that Bowlby described as being rudimentary parts of the attachment control system can be considered as part of the Pavlovian control system. Allowing this system to control behaviour affords speedy response and requires minimal cognitive load. However, its major disadvantage is that it is inflexible, and its responses are only adaptive if you current environment matches that in which its settings were evolved. The avoidant behaviour typical of A-type infants in the Strange Situation, and of adolescents and adults in conflict with romantic partners (19, page 179), may arise from triggering Pavlovian aversive responses (16).

5.2 The Habitual controller

The Habitual controller, in common with the Pavlovian controller, exerts its influence on our behaviour outside of our conscious awareness. As with the Pavlovian Controller, it is a relatively fast system, and also involves minimal cognitive load. However, unlike the Pavlovian controller, it can learn from the results of its previous attempts to gain a reward or avoid a punishment in particular environmental states, and might therefore be compared to the fixed action patterns described by Bowlby. A drawback of the Habitual controllers learning mechanism is that its initial attempts to accomplish a task are often far removed from the optimal performance that it can ultimately gain after extensive experience. Where Model-based control (described in the next section) may include all available information to decide which action to take, habitual control may use just a single simple metric that represents the overall utility of a taking a particular action for a particular state of the environment. This cached value can be seen as comparable to what would be gained if the total anticipated value of all future actions were collapsed to a single value (16). The Habitual Controller has a notable drawback - that the cached values are not open to reflection and cannot be changed in a 'one-shot' manner when the environment changes or when new information comes to light. Since attachment behaviours only change gradually, and people often struggle to reflect on their attachment related actions (19, 43-46), it may be that much of our attachment behaviour is under Habitual control.

According to the Bayesian framework, actions directed by the Pavlovian control system can help shape or interfere with instrumental responses arising from the habitual control system (15). It may be that Pavlovian Instrumental Transfer (PIT) from biases in frequently activated reflex responses to the Habitual Controllers causes the observed continuity in attachment behaviour patterns. Processes that bring controllers into indirect coordination without direct communication may occur throughout the life-span and bring Pavlovian and Habitual controllers into congruent attachment behaviour patterns in a process outside of an individual's awareness.

5.3 The Model-based controller

This controller (also described as the goal-directed controller) involves conscious reasoning in pursuit of explicit goals (14; 18). It can construct plans, representing all its alternative choices and their consequences as models of the self and environment. As it models the environment, this controller can use all available information about the task and environment in its calculations, and when something changes in the environment, it can immediately incorporate this new information in a statistically efficient manner in its model. It therefore has immediate and optimal access to the results of experience. However, this very flexibility leads to its major disadvantages. Considering all available data can be overly time consuming and prohibitive in computational load for all but the simplest problems.

It is likely that Model-based control does not have a central role in controlling infant attachment behaviour of infants, but explicit reasoning about what actions to take within adolescent and adult relationships, including caregiving, seems a reasonable proposal. This might be construed as implementing Bowlby's important aspects of working models concept. It might also prove useful in modeling the goal corrected partnership. However, the reward focused Habitual Controller might also be recognised as a rather minimal form of IWM as it predicts optimal actions from experience.

It was noted above that the observed continuity of attachment behaviour patterns through the life-span might arise when Pavlovian responses shape Habitual behaviour. Might an analogous process occur between the unconscious processing by Pavlovian and Habitual controllers, and attachment behaviour that resulted from the operation of the Model-based controller? The issue here is how does implicit knowledge become explicitly represented. One possibility is that individuals observe and recognise their own unconsciously produced behaviour patterns. They may then elaborate upon these and use them in building their own personal narratives that are congruent with the original behaviour patterns. Thus implicitly representations directing Pavlovian and Habitual patterns of behaviour may not directly turn into explicit representations that are held within the Model-based system. Rather, the unconscious controllers may drive behaviour patterns that are then observed in their interaction with the environment, and it is an individual's interaction with their environment that is observed by their conscious mind, in form of the Model-based controller.

If response patterns arising in the Pavlovian and Habitual controllers may influence response patterns in the Model-based controller, the reverse may also occur. The Model-based system may provide a kind of 'will power' for individuals to overcome their own unconsciously produced responses that they recognise and wish to minimise (15). For example, in adult romantic relationships some individuals (perhaps after advice) may attempt to limit their habitual avoidance of intimacy. Whereas others individuals, who often become emotionally enmeshed, may recognise the benefit of deferring intimacy (19, pages 178-180).

5.4 The Episodic controller

The Episodic controller fits a gap between the Model-based and Habitual controllers (22). This might be when reasoning is too slow, or requires too many cognitive resources, for the Model-based controller to operate effectively, and in addition, the situation that has not been experienced enough times for the Habitual response to be optimised. This controller simply chooses an action that gained a favourable response in a similar previous situation, even if this situation were experienced rarely Recalling the sensitivity of infant behaviour to context when the Strange Situation is repeated within 10-14 days, this controller might therefore be a good candidate explanation for infant behaviour when re-tested in the Strange Situation after only a few weeks. In these cases the infant is in an unusual situation that has only been experienced in this precise manner once before. Since in the original Strange Situation all infants show some level of separation protest, and the carers always returned, it may be an Episodic controller that activates increased separation protest when this procedure is repeated.

5.5 Self-constructing control systems

In addition to explaining how behavioural patterns are transferred between subsystems (which already exist), one of the key future directions for research is how subsystems in the attachment control system actually become constructed. In human subjects it may be that a deliberative system similar to the Model-based controller starts to increasingly direct attachment behaviours from late infancy through to adulthood. As it does so, the form of the IWM's and other representations that it utilises, such as natural language, will become more sophisticated. These higher level representations may then provide a richer substrate for the acquisition of new habitual skills in influencing an individual's attachment relationships (18).

Hierarchical structure in an Habitual controller may thus develop following the employment of hierarchically structured actions by the Model-based controller. However, how does hierarchical structure develop within the Model-based controller? As described in section 3.5, causal hierarchies related to the goal of feeding can develop from actions such as pecking in chicks or sucking in puppies. In these examples, the pecking or sucking actions are initially unrelated to the pursuit of food. From the broad initial range of behaviours (that might be determined in the genome), only a narrower range is selected to be 'set' as feeding behaviours that predictably result in gaining food. So later in development only this subset of the original behaviours is triggered by food stimuli. In altricial species it may be common that particular sets of genome determined behaviours are only recognised post-natally as possessing causal potential. Delaying acquisition of associations between action and goal may be repeated numerous times in a cascaded developmental process (11). This process can then lead to hierarchical action control structures forming as actions and goals are represented at greater levels of granularity (11).

6 Conclusion

This paper has described a range of attachment behavioural phenomena and presented John Bowlby's (6; 7; 8) explanatory framework for these phenomena. It has then surveyed recent autonomous agent simulations which have built upon information processing aspects of Bowlby's framework. These simulations have highlighted limitations in the current understanding of how subsystems within the overall attachment control system may interrelate and become constructed through development. Two AI frameworks, one Bayesian (14) and the other concerned with cascaded hierarchical development (11) have been presented that might inspire developmental explanations of attachment. Key benefits of the Bayesian framework are threefold. In this approach: attachment behavioural phenomena can be mapped onto the performance of constituent subsystems that have been implemented in simulations of other behavioural phenomena; the subsystem interactions in simulations of the attachment control system can be decided in a principled 'Bayesian' manner; and lower level reactive components can 'shape' the nature of patterns of attachment response in higher level deliberative components. Simulating attachment development as a example of a cascaded developmental process may shed light on how the hierarchical structure of the Attachment control system is constructed. Contemporary measurement tools allow attachment response patterns observed in infancy to be compared with those exhibited in adulthood (31). Computational modelling with inspiration from AI may provide new explanations for observed continuities in patterns of attachment response across the life-span.

REFERENCES

[1] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, *Patterns of Attachment: a psychological study of the strange situation*, Erlbaum, Hillsdale, NJ, 1978.

[2] J.W. Anderson, 'Attachment behaviour out of doors', in *Ethological Studies of Child Behaviour*, ed., N. Blurton-Jones, 199–215, Cambridge University Press, Cambridge, UK, (1972).

[3] N. Bischof, 'A systems approach toward the functional connections of attachment and fear', *Child Development*, **48**(4), 1167–1183, (1977).

[4] J. Bowlby, 'Forty-four juvenile thieves: Their character and home life.', *International Journal of Psychoanalysis*, **25**, 1–57, (1944).

[5] J. Bowlby, 'Grief and mourning in infancy and early childhood', *The psychoanalytic study of the child*, **XV**, 9–52, (1960).

[6] J. Bowlby, *Attachment and loss: volume 1 attachment*, Basic books, New York, 1969. (Second edition 1982).

[7] J. Bowlby, *Attachment and loss: volume 2, Separation: Anxiety and Anger*, Basic books, New York, 1973.

[8] J. Bowlby, *Attachment and loss: volume 3 loss, sadness and depression*, Basic books, New York, 1980.

[9] J. Bowlby and J. Robertson, 'A two year old goes to hospital', *Proceedings of the Royal Society of Medicine*, **46**, 425–427, (1952).

[10] I. Bretherton, 'The origins of attachment theory: John bowlby and mary ainsworth', *Developmental Psychology*, **28**(5), 759–775, (1992).

[11] J.M. Chappell and A. Sloman, 'Natural and artificial metaconfigural altricial information-processing systems', *International Journal of Unconventional Computing*, **3(3)**, 211–239, (2007).

[12] N.D. Daw, Y. Niv, and P. Dayan, 'Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control', *Nature Neuroscience*, **8**, 1704–1711, (2005).

[13] P. Dayan, 'The role of value systems in decision making', in 'Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions' *, eds. C. Engel and W. Singer*, 51–70, MIT Press, Frankfurt, (2008).

[14] P. Dayan, 'Goal-directed control and its anitpodes', *Neural Networks*, **22**, 213–219, (2009).

[15] P. Dayan, Y. Niv, B. Seymour, and N.D. Daw, 'The misbehaviour of value and the discipline of the will', *Neural Networks*, **19**, 1153–1160, (2006).

[16] P. Dayan and B. Seymour, 'Values and actions in aversion.',', in 'Neuroeconomics: Decision making and the brain' *, eds. P. Glimcher and C. Camerer and R. Poldrack and E. Fehr*, 175–191, Academic Press, New York, (2008).

[17] D.C. Dennett, *Kinds of minds: towards an understanding of consciousness*, Weidenfeld and Nicholson, London, 1996.

[18] K. Douglas, 'The other you', *New Scientist*, **196**(2632), 42–46, (2007).

[19] S. Goldberg, *Attachment and Development*, Arnold, London, 2000.

[20] 'H. Harlow, 'The Nature of Love', *American Psychologist*, **13**, 573–685, (1958).

[21] J. Holmes, *John Bowlby and Attachment Theory*, Routledge, 1993. (revised edition).

[22] M. Lengyel and P. Dayan, 'Hippocampal contributions to control: The third way', in *21st Annual Conference on Neural Information Processing Systems*, pp. 1–8, (2007).

[23] K. Lorenz, *King Solomons Ring*, Methuen, London, 1952.

[24] P. Meehl, 'Clarifications about taxometric method', *Applied and Preventative Psychology*, **8**, 165–174, (1999).

[25] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).

[26] D. Petters, *Designing Agents to Understand Infants*, Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006. (Available online at http://www.cs.bham.ac.uk/research/cogaff/).

[27] D. Petters and E. Waters, 'Epigenetic development of attachment styles in autonomous agents', in *Proceedings of the Eighth International Conference on Epigenetic Robotics. Modeling Cognitive Development in Robotics Systems*, eds., M. Schlesinger, L. Berthouze, and C. Balkenius, 153–154, Lund University Cognitive Studies, 139, (2008).

[28] L.A. Sroufe, 'Attachment classification from the perspective of infant-caregiver relationships and infant temperament.', *Child Development*, **56**(1), 1–14, (1985).

[29] E. Waters, K. Kondo-Ikemura, G. Posada, and J. Richters, 'Learning to love: Mechanisms and milestones', in Minnesota Symposium on Child Psychology (Vol. 23: Self Processes and Development)*, eds. M. Gunner & Alan Sroufe*, 217–255, Psychology Press, Florence, KY, (1991).

[30] E. Waters, S. Merrick, D. Treboux, J. Crowell, and L. Albersheim, 'Attachment stability in infancy and early adulthood: A 20-year longitudinal study', *Child Development*, **71**, 684–689, (2000).

[31] H. Waters and E. Waters, 'The attachment working models concept: Among other things we build script-like representations of secure base experiences', *Attachment and Human Development*, **8**(3), 185–197, (2006).

[32] M. Wooldridge, *An introduction to multiagent systems*, Wiley, chichester, 2002.